关联规则在空间数据挖掘中的应用及实现

张建峰,王 泳,王 剑
(江西理工大学 信息工程学院,江西 赣州 341000)

摘 要:空间数据挖掘是从空间数据库中抽取隐含知识、空间关系及空间数据库中存储的其它信息的方法。空间关联规则是空间数据挖掘的一个重要研究领域,利用空间关联规则把空间数据库中的数据转化为知识是一个很好的方法。在分析空间关联规则的基础上,用基于关联规则的逐步求精挖掘算法,得出空间数据库中的隐含知识,通过实例证明其方法的可行性。

关键词:空间数据挖掘;关联规则;ArcObjects
中图分类号:TP311.13 **文献标识码**:A **文章编号**:1673-629X(2007)08-0208-04

Application and Realization of Association Rules in Spatial Data Mining

ZHANG Jian-feng, WANG Yong, WANG Jian
(School of Information Engineering, Jiangxi University of Science & Technology, Ganzhou 341000, China)

Abstract: *Spatial data mining is a method that gets connotative knowledge, spatial relationship and other information in spatial database.* Spatial association rule discovery in spatial database is a very important reserch field. A spatial association rule is a rule indicating certain association relationship among a set of spatial predicates. In this paper, based on analyzing spatial association rules, an efficient method for mining spatial association rules with spatial analysis in geographic information databases is proposed.

Key words: spatial data mining; association rules; ArcObjects

0 引 言

随着地理信息系统在各个领域的广泛应用,使得专门管理空间数据的数据库系统——空间数据库系统得到前所未有的广泛使用。近年来随着其它空间信息技术如遥感、测绘和GPS技术的飞速发展,使得各种空间数据呈指数级增长。但当面对越来越多的迅速膨胀的数据时,却出现了"数据丰富,知识贫乏"的现象。人们无从着手去理解数据库中包含的信息,更难以获得有价值的信息! 空间数据挖掘技术的出现,为人类充分利用海量的空间数据提供了方法。空间数据挖掘能够从空间数据中提取隐含的信息,找出最有价值的知识来指导决策,已经成为国际上研究和应用的热点。

1 基本概念

1.1 空间数据挖掘

空间数据挖掘(Spatial Data Mining, SDM),或称"从空间数据库中发现知识"(Knowledge Discovery from Spatial Data-bases),是指从空间数据库中提取用户感兴趣的空间模式与特征、空间与非空间数据的普遍关系及其它隐含在空间数据库中的普遍的数据特征。可用于对空间数据的理解、空间关系和空间与非空间数据间关系的发现、空间知识库的构造、空间数据库的重组和空间查询的优化。在地理信息系统、遥感、图像数据库探测、医学图像处理、导航、交通控制、环境研究以及许多使用空间数据的领域中有广泛的应用。

空间数据挖掘的对象是空间数据库,空间数据库是空间数据集合,它实现对具有一定地理要素特征的相关空间数据集合的统一管理,空间数据间紧密联系共同反映现实世界中某一区域内综合信息或专题信息间的联系,主要应用于地理空间数据处理和分析。空间数据库管理的是空间数据对象,不同于一般数据对象。空间数据对象是具有几何属性约束的数据对象。现实世界中的实体都是处于一定的时空中,抽象为信息世界中的数据对象应该具有一般属性、时间属性、几何属性(Geometry包括位置和形状)和行为等特性,然而一般数据库管理的是不包含几何特性的数据对象[1],所以具有以下特点:

收稿日期:2006-11-02
基金项目:国家自然科学基金资助项目(40401045)
作者简介:张建峰(1980-),男,山西原平人,硕士研究生,研究方向为GIS应用开发;导师:兰小机,博士,教授,硕士生导师,研究方向为GIS空间数据互操作研究。

万方数据

(1)数据源十分丰富,数据量非常庞大,数据类型多,存取方法复杂;

(2)应用领域十分广泛,只要与空间位置相关的数据,都可以对其进行挖掘;

(3)挖掘方法和算法非常多,而且大多数算法比较复杂,难度大;

(4)知识的表达方式多样,对知识的理解和评价依赖于人对客观世界的认知程度。

目前,空间数据的组织形式一般为图层。多个具有某些相同或相似特性的空间对象的集合,在数据库中以表的形式进行组织和表达,为了提高地图中各个要素的检索速度,便于数据的灵活调用、更新及管理,在空间数据库中,往往将不同类、不同级的图元要素进行分层存放,每一层存放一种专题或一类信息。按照用户一定的需要或标准把某些相关图元要素组合在一起成为图层,它表示地理特征以及描述这些特征的属性的逻辑意义上的集合。这样一个综合地图可包括许多层图层的叠加,我们所要做的任务就是在这样的多图层数据集中,挖掘我们所关心的知识。

1.2　关联规则

关联规则[2]是表示数据库中一组对象之间某种关联关系的规则,关联规则用数学模型描述如下:设 $I=\{i_1,i_2,\cdots,i_m\}$ 是 m 个不同项目的集合,设任务相关的数据 D 是数据库事务的集合,其中每个事务 T 是项的集合,使得 $T\subseteq I$,每个事务有一个标识符,称为 TID。关联规则表示为蕴涵式 $X\Rightarrow Y$,其中 $X\subset I$, $Y\subset I$, $X\cap Y=\varnothing$。交易集 D 中的规则 $X\Rightarrow Y$ 由置信度 C(Confidence) 和支持度 S(Support) 约束。置信度 C 定义为:D 中含有 X 中的交易的 $C\%$ 也含有 Y。支持度 S 定义为:包含 $X\cup Y$ 的交易占 D 的 $S\%$。置信度表示蕴含式的强度,支持度表示在规则中出现该型式的频度,具有高置信度和高支持度的关联规则称为强关联规则。

在实际研究中,满足一定的置信度和支持度的关联规则才有意义。为此需要定义两个阈值:最小置信度(记为 Min_ Conf) 和最小支持度(记为 Min_ Sup)。如果项目集 $X\subset I$, $\sup(X\Rightarrow Y)\geqslant$ Min_ Sup,则称 X 是频繁项集;如果 $\mathrm{conf}(X\Rightarrow Y)\geqslant$ Min_ Conf,则称规则 $X\Rightarrow Y$ 成立。关联规则的挖掘是在事务数据库 D 中找出具有用户给定的最小支持度 Min_ Sup 和最小置信度 Min_ Conf 的关联规则。

1.3　Apriori 算法介绍

Apriori 算法是一种最有影响的挖掘布尔关联规则频繁项集的算法,算法使用频繁项集性质的先验知识,用逐层搜索的迭代方法来获得频繁项集。K_ 项集用于探索 $(k+1)$ 项集。首先找到频繁 1 项集的集合,记为 L_1。L_1 用于找频繁 2 项集的集合 L_2,如此下去,直到不能找到频繁 K 项集[3]。

Apriori 性质:

频繁项集的所有非空自己都必须也是频繁的。这是因为根据定义,假设项集 I 不满足最小支持度(min_ sup),则 I 不是频繁的,如果把项 A 添加到 I,则结果项集(即 IUA) 不可能比 I 更频繁出现。因此,结果项集也不是频繁的。

利用 Apriori 性质,通过连接和剪枝两步过程来实现频繁项集的挖掘。

a. 连接步:

通过对 L_{k-1} 中的每个元素执行连接,得到了 L_k 的候选集合 C_k。

b. 剪枝步:

C_k 是 L_k 的超集,即它的成员中也有不是频繁的。首先,根据 Apriori 性质,缩小 C_k 的范围,然后扫描数据库,确定 C_k 中每个候选的计数,从而确定 L_k。

2　空间关联分析

2.1　空间关联规则

在自然和人文界中,各种地理要素(现象)的分布并不是孤立的。它们相互影响,相互制约,彼此之间存在着一定的联系。空间关联规则,也称空间相关关系,主要指空间对象之间的空间和非空间关系,空间关联知识对于智能化的空间分析或空间辅助决策支持来说是十分重要的。因此,从空间数据库中发现空间关联规则是空间数据挖掘的重要任务之一。空间关联规则挖掘的目的是发现现实世界中空间对象之间的有趣的关联模式或相互关系。

空间关联规则的形式有多种,如空间目标之间相离、相邻、相连、共生、包含、被包含、覆盖、被覆盖、交叠等空间拓扑关系是典型的空间关联规则;“居民地(城镇)与道路相连”,“道路与河流的交叉口是桥梁”等用自然语言描述的知识同样属于规则范畴。空间分布规律本质上也是一种空间关联知识,它反映了所感兴趣的对象与空间位置或高程的关联,等等。

目前研究最多的一种空间关联规则形式为 $A\Rightarrow B[S\%,C\%]$,其中 A 和 B 是空间和非空间谓词的集合,$S\%$ 表示规则的支持度,$C\%$ 表示规则的可信度,这种形式的空间关联规则的定义与相关算法是直接从事务型数据库的关联规则挖掘延伸过来的。由于空间相关关系的内涵十分丰富,包括了空间对象之间的各种各样的内容联系,因此,形如 $A\Rightarrow B[S\%,C\%]$ 的空间关联规则是对空间关联规则的一种较狭义的定

万方数据

义,并且它也是比较有代表性的空间关联规则之一。

空间关联规则的一般形式是:

$A_1 \wedge A_2 \wedge \cdots \wedge A_m \Rightarrow B_1 \wedge B_2 \wedge \cdots \wedge B_n$

谓词 $A_1, A_2, \cdots, A_m, B_1, B_2, \cdots, B_n$ 是空间和非空间谓词的集合,其中至少有一个是空间谓词;令 $A = A_1 \wedge A_2 \wedge \cdots \wedge A_m$,称为规则的前件;令 $B = B_1 \wedge B_2 \wedge \cdots \wedge B_n$,称为规则的后件;$A \wedge B = \varnothing$;$S\%$ 是规则的支持度(support),$C\%$ 表示规则的置信度(confidence)[4]。例如,下面是一个空间关联规则的例子:

is_ a(X, "school") ∧ Close_ to(X, "sports_ center")⇒Close_ to(X,"park")[3%,70%]

此规则表明 70% 靠近体育中心的学校也靠近公园,并且有 3% 的数据符合这一规则。

置信度是对关联规则的准确度的衡量,支持度是对关联规则的重要性的衡量。支持度说明了这条规则在所有空间对象中有多大的代表性,显然,支持度越大,关联规则越重要。有些关联规则置信度虽然很高,但支持度却很低,说明该关联规则实用机会很小,因此并非很重要。

在该例中 Close_ to 是一种空间谓词,Close_ to(A, B) 表示 A 和 B 接近,Close_ to 谓词定义在知识库中,此外还有其它的空间谓词可以用来构成空间关联规则,如代表空间方位的:Left_ of(左边),West_ of(西部),North_ of(北边),代表拓扑关系的:Intersect(相交),Overlap(重叠),代表距离关系的:Close_ to(临近) Far_ away(远离)等等。

2.2 空间挖掘步骤

R. Agrawal 等人首先提出了关联规则的采掘问题并给出了解决此问题最原始的算法 AIS[3] 之后,该问题得到了国际人工智能和数据库等领域学者的密切关注,提出了多种算法。所有的挖掘算法不论采用什么样的数据结构,其复杂程度,效率如何,它们实现步骤大致相同,空间关联规则也不例外,主要有以下几个步骤:

1)对空间信息进行抽象得到抽象数据,再经过预处理得到挖掘任务有用的数据。根据具体问题的要求对空间数据库进行相应的操作,从而构成规格化的数据库 D;

2) 针对 D,求出所有满足最小支持度的项集,即大项集。由于一般情况下所面临的数据库都比较大,所以此步是算法的核心;

3) 生成满足最小置信度的规则,形成规则几何 R;

4) 解释并输出 R。

2.3 空间挖掘算法

由于空间关联的挖掘[5]需要在大量的空间对象中计算多种空间关系,因此其代价是很高的。一种称为逐步求精的挖掘算法可用于空间的关联分析,该方法首先用一种快速的算法粗略地对一个较大的数据集进行一次挖掘,然后在裁剪过的数据集上,用代价较高的算法进一步改进挖掘的质量。

逐步求精的算法可描述为:

1)进行空间分析,在一定距离内(缓冲区分析)内得到挖掘的目标;

2)采用 MBR 技术,R+ 树等索引方法进行快速查询;

3)由前一步生成的空间分析的拓扑关系数据表,计算这些谓词的支持度,滤去支持度小的项目;

4)采用 MBR 技术对第三步裁忮后的空间谓词关系进行检查,滤去与实际不相符的空间谓词关系,形成新的拓扑关系数据表,并计算这些谓词的支持度,滤去支持度最小的项目。

3 实例分析

3.1 ArcObjects 简介

ArcObjects[6]是 ESRI 公司 ArcGIS 家族中应用程序 ArcMap, ArcCatalog 和 ArcScene 的开发平台,它是基于 Microsoft COM 技术所构建的一系列 COM 组件产品。ArcObjects 不是为最终用户而是专门为开发人员提供的二次开发软件,通过 ArcObjects,用户可以非常方便地开发出功能强大的 GIS 应用系统。

ArcObjects 是一个组件产品,它可以用于大量开发框架中,包括流行的像 Visual C++, Visual Basic, Delpi, .NET 等程序设计环境,因此开发人员可以在自己熟悉的开发环境中利用 ArcObjects 开发 GIS 应用。

ArcObjects 支持多种空间数据格式,如:ArcInfo COVERAGE, ESRI Shape files, Geodatabase, DXF, DWG 等。提供了丰富的空间分析组件,通过这些组件,用户可以实现诸如空间查询、缓冲区分析、叠置分析和网络分析等功能。

3.2 程序实现

根据以上分析,通过 ESRI 公司的 ArcEngine 平台,开发语言为 C#,在某市的基础地理信息系统实现了一个基于关联规则的空间数据挖掘模块。该模块的目的是在大量的空间数据中,找出学校周围各种商店的情况。

空间谓词 Close_ to(A,B)计算如下:

万方数据

```
AxesriMapControl. AxMapControl axMapControl1;
axMapControl1 = new AxesriMapControl. AxMapControl();
axMapControl1. LoadMxFile(@"C: \ City. mxd",0,null);
IMap pMap = axMapControl1. Map;
IFeatureLayer pSchoolFeatureLayer = pMap. get_ Layer(0);
IFeatureLayer pShopFeatureLayer = pMap. get_ Layer(1);
//通过点选获得学校要素
IFeature pSchoolFeature;
//通过缓冲区操作获得学校周围的商店
IPoint pPoint = pSchoolFeature. Shape;
ITopologicalOperator pTopo = pPoint as ITopologicalOperator;
IGeometry pBuffer = pTopo. Buffer(100);
IGeometry pGeometry = pBuffer. Envelope;
ISpatialFilter pSpatialFilter;
pSpatialFilter. Geometry = pGeometry;
pSpatialFilter. SpatialRel = esriSpatialRelEnum. esriSpatial-
RelWithin;
IFeatureSelection pFeatureSelection;
pFeatureSelection. SelectFeatures ( pSpatialFilter, esriSelection-
ResultEnum. esriSelectionResultNe - w,false);
```

esriSpatialRelEnum. esriSpatialRelWithin 是 ArcObjects 提供的一个常量,它选择那些落入目标几何图形内的要素。通过对选择集应用 Apriori 算法,由此就可以从大量的数据文件中获取我们想要的知识。

4 结　语

讨论了空间数据关联规则挖掘的算法,同时给出了一个实例。随着计算机在各个领域的应用,使得空间数据库系统的应用越来越广泛。空间数据挖掘将人工智能、机器学习、可视化技术、数据库等集于一体,是一门新兴的交叉学科;空间数据挖掘不同于一般的数据挖掘,这是由于空间数据挖掘的对象不仅包括空间属性数据同时还包括非空间属性数据,但是两者之间又存在着一定的联系,进一步将一般的数据挖掘技术应用到空间数据挖掘领域,还需要更深的研究;另外,除关联规则外,空间聚类规则、空间分类规则、空间分布规律等也是空间知识发现的重要方法。因此,借助计算机进行自动的分析数据抽取知识,是一个重要的研究方向。

参考文献:

[1] 倪　凯,祝晓东,张　超.基于关联规则的空间数据知识发现及实现[J].计算机应用与软件,2005(12):34-35.

[2] Han Jiawei, Kamber M. 数据挖掘:概念和技术[M].范明,孟小峰,等译.北京:机械工业出版社,2001.

[3] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Item in large Databases[C]// In: Proceedings of 1993 ACM - SIGMOD International Conference Management of Data (SIGMOD'93). Washington D. C.: [s. n.], 1993:207-216.

[4] 李国锋. 空间数据挖掘技术研究[D]. 西安:西安电子科技大学,2005.

[5] Agrawal R, Srikant R. Fast algorithms for mining association rules[C]// In: Proc of 20th Int Conf Very Large Database. (VLDB'94). CA:[s. n.],1994:487-499.

[6] 韩　鹏,徐占华,褚海峰,等.地理信息系统开发—ArcObjects 方法[M].武汉:武汉大学出版社,2005.

(上接第 207 页)

```
wfengine.Dispose(); //销毁工作流
}
    }
    else
    {
    ReportErro(strEngineModal,strOperationNo);//报告错误,表示调用错误
        }
    }
    else
    { ReportErro(strEngineModal);//返回错误,表示工作流模型不存在}}
```

3 结束语

介绍了科学技术奖励评审平台的设计与实现,该系统采用多层结构,并从系统需要支持自动化与半自动化的业务流程,引入了工作流技术。使得该平台能够快速地实现对业务的重组。目前,该平台已经在国家科技奖励办公室以及上海、山东、内蒙古、湖南、河北、甘肃、浙江等省、市、自治区级科技奖励办公室投入使用,深受用户的好评。

参考文献:

[1] 张友生.系统分析与设计技术[M].北京:清华大学出版社,2005.

[2] 罗海滨,范玉顺,吴　澄.工作流技术综述[J].软件学报,2000,11(7):899-907.

[3] Workflow Management Coalition. Workflow management coalition terminology and glossary[R]. WfMC - TC - 1011. Brussels: Workflow Management Coalition, 1996.

[4] 范玉顺.工作流管理技术基础[M].北京:清华大学出版社,2001.

[5] 饶　元,冯博琴,李尊朝.基于 WebServices 的服务合成技术研究综述[J].系统工程与电子技术,2005,27(8):1481-1489.

万方数据